ANALYSIS AND OPTIMIZATION OF CYCLIC METHODS IN ORBIT COMPUTATION

TR-05-071-005-1

February 1973

Prepared By

S. Pierce

Department of Mathematics

Computer Science Council

California State University, Fullerton

for

Trajectory Analysis & Geodynamics Division

Goddard Space Flight Center

National Aeronautics & Space Administration

TABLE OF CONTENTS

1. INTRODUCTION

This is a semi-annual Status Report for the work accomplished under NASA Grant NGR-05-071-005 for the Trajectory Analysis and Geodynamics Division, Goddard Space Flight Center, National Aeronautics and Space Administration.

This report contains the mathematical analysis and computation of the K=3, order 4; K=4, order 6; and K=5, order 7 cyclic methods and the K=5, order 6 Cowell method and some results of "optimizing" the 3 backpoint cyclic multistep methods for solving ordinary differential equations [2,5,10] . Cyclic methods have the advantage over traditional methods of having higher order for a given number of backpoints while at the same time having more free parameters. After considering several error sources the primary source for the cyclic methods has been isolated.

The free parameters for three backpoint methods were used to minimize the effects of some of these error sources. They now yield more accuracy with the same computing time as Cowell's method on selected problems.

This work is being extended to the five backpoint methods. The analysis and optimization are more difficult here since the matrices are larger and the dimension of the "optimizing space" is larger. Indications are that the primary error source can be reduced. This will still leave several parameters free to minimize other sources.

2. COMPUTATION RESULTS

Class II Methods on Orbits (Local Correction Error):

Integrating the two body equations (6.1) for the motion of a satellite similar to GEOSB with exact starting values in [5] we found that with the cyclic K=5, order 7 corrector and the PECE algorithm moving from an unstable order 7 predictor to Störmer order 6 predictor increased the error slightly at small h (due to lower order) but greatly decreased the error at larger h (due to increased stability). This implies that the predictor has a great effect on the cyclic method and should be included in future derivations.

With the latter predictor the $PE(CE)^2$ algorithm improved the accuracy slightly at small h and greatly at larger h. However, iterating to convergence with the corrector at each step did not decrease the errors any more than this.

As can be seen from Figure 1 the cyclic K=5 corrector still needs improvement as it is being compared to the cyclic K=4 and Cowell K=5 with only the PECE algorithm. At this h=1 sec. the "random" local roundoff error probably dominates the local truncation error. The cyclic K=4 error typically oscillates in the first few steps while the K=5 jumps greatly. Details of Cowell on this equation were not available but other results indicate it increases slowly. All three soon level off to about the same rate of error growth.

Class I Methods on $y' = .5y$:

Because of the strong dependence on predictor and algorithm and because the error curves were similar to those for orbits, we decided to being our study on linear equations. Figure 2 compares the cyclic (order 9) and Adams (order 9) class I, K=5 methods using program COMPAR in double precision with exact starting values. Since 9x4=36 and since the CDC machine carries $\sim$28 places the "random" local roundoff error dominates the local truncation error. The $h=10^{-6}$ curves are similar to these $h=10^{-4}$ curves.

The Adams coefficients were "perturbed" by 5×10^{-26} (see Section 3.1) which is the estimated error in the cyclic coefficients. This is probably the reason the Adams error greatly increases during the first cycle.

Class II Methods on $y'' = y$:

In order to dominate the truncation and roundoff ($\sim10^{-28}$) errors the starting error $=(1,-2,4,-8,16) \times 10^{-20}$ was propogated using COMPAR at h=0 in double precision for the cyclic, K=4, order 6; cyclic, K=5, order 7; and Cowell, K=5, order 6 methods. See Figure 3.

Figure 4 shows the results of integrating $y'' = y$ with COMPAR in double precision at $h=10^{-6}$ with exact starting values. Since 6x6=36 the roundoff error dominates.

In Figure 5 at $h=10^{-2}$ with exact starting values the truncation error dominates. The cyclic method is better in the first cycle since its order is higher but immediately gets worse in the second.

The cyclic, K=5, order 7-4/5 curves were almost the same as the order 7 ones so they were not graphed. Single and double precision runs were made with h ranging from 0 to 10^{-1}. All methods blew up at 10^{-1} (i.e., error after 50 cycles ~ 1) but remained good at $h=10^{-2}$.

3. COMPUTATIONAL ANALYSIS

3.1 Stability With Respect to Perturbations of the Coefficients:

Because of the propogation of roundoff errors in solving the nonlinear stability equations, the K=5 cyclic coefficients were estimated to be in error by $\sim 10^{-25}$. The K=6 even more so because of the inaccurate order equation solutions. These errors do not effect integrations in single precision (~ 14 places) so the conclusions will still hold since they will be based in part on single precision integrations. However, the sensitivity of the integration errors in double precision to a perturbation of the coefficients is a measure of the stability of a method.

Table 1 shows the effects of adding to each cyclic, K=5, order 7 coefficient $5\text{x}10^{-23}$ and to the Cowell, K=5, order 6 ones the same amount with exact starting values at $h=10^{-4}$ where both roundoff and truncation contribute. The Cowell errors increase by a factor of 10^7 and the cyclic by 10^5 (less because they were in error to start with). Since the machine computes with 28 to 29 places our perturbation is $\sim 10^5$ times the computation accuracy, the computations will now be done 10^5 times less accurate, and this is about the error increase shown. If this result can be generalized then the methods are nearly equally sensitive but not overly sensitive to these perturbations. This implies derivations of coefficients must be done more accurately than is their practical use in integrating equations.

Table 1. Integration Errors

Error at Cycle #	#1	#2	#50
Cyclic unperturbed	2 E-24	6 E-23	2 E-20
Cyclic perturbed	4 E-19	2 E-17	2 E-14
Cowell unperturbed	2 E-27	7 E-27	4 E-24
Cowell perturbed	2 E-20	6 E-20	3 E-17

3.2 Local Truncation Error (Order):

The order equations were satisfied in all cases and the first few non-zero coefficients in the local truncation error expansions are comparable to Cowell's.

Table 2. Truncation Error Coefficients

Cyclic, K=4, order 6

Method #	=	1,3	2,4
C_8	=	.0021	.0026
C_9	=	.0042	.0052
C_{10}	=	.0046	.0056
C_{11}	=	.0035	.0042

Cyclic, K=5, order 7

Method #	=	1	2	3	4	5
C_9	=	-.0005	.0026	-.0003	-.0007	-.0009
C_{10}	=	-.0011	.0064	-.0007	-.0017	-.0022
C_{11}	=	-.0012	.0085	-.0007	-.0019	-.0026
C_{12}	=	-.0009	.0079	-.0003	-.0014	-.0021

3.3 Eigenvalues, Vectors, Condition Numbers (Stability):

The "stability matrix" determines how the local errors are propogated for the cyclic methods and for a traditional method used cyclicly. The eigenvalues were of primary concern and were computed for $h=0$ and for $h=10^{-6}$, 10^{-4}, 10^{-2}, 10^{-1} for the equation $y'' = y$. The $h=0$ computations were not accurate since

the Cowell and cyclic matrices were ill-conditioned, the Jordon block associated with $\lambda=1$ was 2x2, and the computed eigenvalues were overly sensitive to round-off errors. The principal condition numbers as derived from the eigenvectors (Section 4) were also of importance and were computed at $h=10^{-4}$ (Section 5.1 contains more thorough computations in this regard). Finally the row norms of the stability matrices varied slowly with h and are: 9.0 for the cyclic K=4 method, 11.0 for the Cowell K=5 method, and 1723.8 for the cyclic K=5 method.

The cyclic and Cowell K=5 principal eigenvalues behave the same. The Cowell extraneous values remain well within the unit circle at $h=10^{-1}$ when it blows up in computations implying it is "over stable." The cyclic extraneous values leave the unit circle at the same h at which it blows up implying that the "blow up" point may be moved to larger h by improving the behavior of the eigenvalues.

Table 3. Eigenvalues, λ_k

In the vector (r,i) r is the real part, i the imaginary.

h	Root #1	#2	#3	#4	#5
			Cyclic K=4		
10^{-6}	1.000004	0.999996	(-1.0,1.4E-13)	(-1.0,-1.4E-13)	
10^{-4}	1.00040	0.99960	(-1.0,1.4E-9)	(-1.0,-1.4E-9)	
10^{-2}	1.04	0.96	(-1.0,1.4E-5)	(-1.0,-1.4E-5)	
10^{-1}	1.5	0.67	(-1.0,1.4E-3)	(-1.0,-1.4E-3)	
			Cowell K=5		
0	1.	1.	-5.E-23	-7.E-35	-3.E-35
10^{-6}	1.000005	0.999995	1.E-28	-8.E-28	-1.E-28
10^{-4}	1.00050	0.99950	5.E-18	(-2.,-4.)E-18	(-2.,4.)E-18
10^{-2}	1.05	0.95	2.E-11	(-1.,-2.)E-11	(-1.,2.)E-11
10^{-1}	1.65	0.61	4.E-8	(-3.,-4.)E-8	(-3.,4.)E-8
			Cyclic K=5		
0	1.	1.	-1.E-4	(0.7,1.)E-4	(0.7,-1.)E-4
10^{-6}	1.000005	0.999995	2.E-6	(-1.,2.)E-6	(-1.,-2.)E-6
10^{-4}	1.00050	0.99950	0.00010	-1.2E-5	-9.1E-5
10^{-2}	1.05	0.95	0.02	-1.E-5	-0.005
10^{-1}	1.65	0.61	1.56	-1.E-5	-0.006

Table 4. Principal Condition Numbers, s_k

Root #	#1	#2	#3	#4
Cyclic K=4	-1.414E-4	+1.414E-4	(-0.4,1.2)	(-0.4,-1.2)
Cowell K=5	6.326E-5	-6.324E-5		
Cyclic K=5	-1.033E-8	+1.032E-8		

4. MATHEMATICAL ANALYSIS AND OPTIMIZATION

4.1 Summary of Computational Results:

The Class I and II, cyclic, K=5 methods in computations dominated by starting error, "random" roundoff error, or truncation error always show an error jump in the first or second cycles. The error growth levels off to the same rate as Cowell's as you integrate along (increase cycle number) on a given equation so this is not a stability problem (Figures 1,2,3,4 and Table 3). The difference between the cyclic and Cowell errors remains a factor of about 5×10^3 for $h=0,10^{-6}$ and 10^{-4} (compare Figures 3,4 and Table 1). At 10^{-2} (Figure 5) the truncation error dominates the first cycle so the cyclic error is smaller but during the second cycle it jumps as before but by a smaller factor of about 5×10^2.

In fact it is quite common that the cyclic error actually decreases (figures 2 and 3) while this has never been observed for any traditional method. This is probably due to the different correctors cancelling the errors. It may be possible to choose the correctors so that the errors exactly cancel within each cycle thus obtaining an "errorless" method. This does not seem to be possible with any single method. The K=4 cyclic method illustrates something very close to this "ideal" cyclic method. The error growth in the first few cycles is smaller and thereafter levels off at the same rate as Cowell but at a smaller error level.

The cyclic K=4 method has the largest condition numbers, the smallest norm, the slowest growth of extraneous values with h, and the smallest error in all computations. The cyclic K=5 is just the opposite in all respects. The Cowell K=5 is in-between in all respects its condition numbers being about 5×10^3 larger than the cyclic at $h=10^{-4}$ (Table 4).

4.2 Convergence Proof, Error Bound:

Some analysis will be provided to explain these results and to lead to methods of improving the accuracy. The equation $y'' = \alpha y$ will be studied in detail and it is expected that improvements will be obtained also on orbit problems.

For $y'' = f(x, y)$ the cyclic methods take the form

(4.2.1) $A_1 Y_{s+1} + A_0 Y_s - h^2 (B_1 Y''_{s+1} + B_0 Y_s'') = 0$ where $s = 1, 2, \ldots, S$ is the cycle number, A_1, B_1 are lower triangular and A_0, B_0 upper triangular $K \times K$ matrices consisting of the a_i and b_i of the K correctors [2], Y_s consists of the approximate solution at the K grid points of the $s^{\underline{th}}$ cycle, Y_s'' are the corresponding approximate second derivatives from $f(x_n, y_n)$, and Y_0 are the starting values.

If each individual method is applied to the exact solution, $y(x)$, restricted to the grid points, we obtain

(4.2.2) $\sum_{k=0}^{K} [a_k^m y(x_{n+k}) - h^2 b_k^m y''(x_{n+k})] = t_m(x_n)$

where $m = 1, 2, \ldots, K$ = the number of the method, the local truncation error $t_m(x_n) = C_0^m y(x_n) + C_1^m y'(x_n)\, h + \cdots + C_p^m y^{(p)}(x_n) h^p + \cdots$ [9, p. 296]. The order of the local truncation error is the power of h in the first non-zero term. The "order" of method #m is the order of $t_m - 2$.

Let $T_{s+1} = [t_1(x_{sK}), t_2(x_{sK+1}), \cdots, t_K(x_{sK+K-1})]$ consist of the truncation errors of the K methods in the $s+1^{st}$ cycle. Since the computer has only finite word length and since we iterate only a finite number of times in solving the implicit corrector equation

at each step, a roundoff error will be committed. The right side of (4.2.2) should be $t_m(x_n) + r_m$ where r_m must be considered random variables in practice. Let $R_{s+1} = [r_1, r_2, \cdots, r_K]$. Writing (4.2.2) in matrix form and subtracting (4.2.1) we obtain

(4.2.3) $A_1 E_{s+1} + A_0 E_s - h^2(B_1 E''_{s+1} + B_0 E''_s) = T_{s+1} + R_{s+1}$

where $E_s = [y(x_{sK}) - y_{sK}, \cdots, y(x_{sK+K-1}) - y_{sK+K-1}]$ and E''_s consists of the exact minus approximate second derivatives.

For $y'' = \alpha y$ (4.2.3) becomes

(4.2.4) $L E_{s+1} + U E_s = T_{s+1} + R_{s+1}$

where $L = A_1 - \alpha h^2 B_1$ and $U = A_0 - \alpha h^2 B_0$ or

(4.2.5) $E_{s+1} = A E_s + B_{s+1}$ where $B_{s+1} = L^{-1}(T_{s+1} + R_{s+1})$ is the total local error magnified by L^{-1} and $A = -L^{-1}U$ is the "stability matrix". In terms of the starting errors, E_0;

(4.2.6) $E_{S+1} = A^{S+1} E_0 + \sum_{s=0}^{S} A^s B_{S-s+1}$. In terms of the Jordan cononical form, J, and the similarity transform, P,

(4.2.7) $E_{S+1} = P J^{S+1} P^{-1} E_0 + \sum_{s=0}^{S} P J^s P^{-1} B_{S-s+1}$.

(4.2.8) Thm: Convergence for $y'' = \alpha y$:

If (i) stability: λ_k are distinct for $h > 0$ and $|\lambda|_{max} = 1 + \epsilon$

where $\epsilon \leq c h$,

(ii) consistency: the local truncation and roundoff errors $||B_s|| \leq O(h^2)$,

then (iii) the cyclic method converges and the order of the propogated error is the min of the orders of the starting error, local roundoff error minus 1 or local truncation error minus 1, and

(iv) $||E_{S+1}|| \leq K^{\frac{3}{2}} |p_i|_{max} [||E_0|| + ||B_i||_{max} (x-x_0)/Kh] e^{c(x-x_0)/K}$

where the terms are explained below.

Proof: Letting X_k be the normalized eigenvector of A associated with λ_k, Z_k that of A^T, $s_k = X_k^T Z_k$ are the "condition numbers", and $p_k = 1/s_k$ then the columns of P are X_k, $||P|| < K$, the rows of P^{-1} are $p_k Z_k$ [12], the rows of $J^s P^{-1}$ are $p_k \lambda_k^s Z_k$, so $||J^s P^{-1}||$ $\leq \sqrt{K}\,|p_k \lambda_k^s|_{max} \leq \sqrt{K}\,|p_k|_{max}(1+\epsilon)^s$. Using inequalities for the row norm in (4.2.7)

$$||E_{S+1}|| \leq ||P||\;||P^{-1}||\,[\,||E_0||\,(1+\epsilon)^{S+1} + ||B_s||_{max} \sum_{s=0}^{S} (1+\epsilon)^s\,]$$

$$\leq K^{3/2}\,|p_k|_{max}\,[\,||E_0||\,(1+ch)^{S+1} + ||B_s||_{max}(S+1)\,(1+ch)^s\,]$$

$$\leq K^{3/2}\,|p_k|_{max}(1+ch)^{(x-x_0)/Kh}\,[\,||E_0|| + (S+1)\,||B_s||_{max}\,]$$

$$\leq K^{3/2}\,|p_k|_{max}\,e^{c(x-x_0)/K}[\,||E_0|| + ||B_s||_{max}(x-x_0)/K\,h\,].$$

In the last steps we used $(1+\epsilon)^S \leq e^{S\epsilon}$ and the fact that if x is in the $S^{\underline{th}}$ cycle then $K(S-1)\,h \leq x - x_0 < K\,Sh$. Since $||B_s|| \leq O(h^2)$ the second term is $O(h)$ and if the starting errors are at least $O(h)$ then we have convergence. End of Proof.

The stability condition is suggested by Table 3 for $\alpha > 0$ but has not been established yet for $\alpha < 0$. It also suggests $c = K$. This convergence proof can be extended to a more general class of $f(x, y)$ by using stability at $h = 0$ (not $h > 0$) and incorporating $h \neq 0$ by using the linearization of $f(x, y)$ given by a Lipschitz condition. The theorem would resemble that hypothesized in [2] with constants resembling those of (4.2.8 iv). Because we are considering linear

equations this bound is better than that in [9] for arbitrary f and traditional methods in several respects. For example, the order of convergence is one higher. Also, for $\alpha > 0$ $y \sim e^x$ so the relative error grows only linearly with x (or with S) as is observed in the graphs.

Both bounds also show if the same local truncation error can be obtained with smaller K then the propogated error will be smaller. This is one advantage of circumventing the Dahlquist [3] stability criteria in addition to that of easier restarting. An advantage of treating the cyclic methods in this matrix fashion instead of as an "auxiliary method" [4] is that the factor $|p_k|_{max}$ is explicit. This is also in agreement with the observations (inc. elliptical orbits) and with Table 4 "explains" why methods differ so greatly in the first cycle but all level off to about the same error growth when the starting and roundoff errors dominate. When the truncation dominates the method with the smaller truncation error will be best in the first cycle since $E_1 = AE_0 + L^{-1}(R_1 + T_1) \approx L^{-1}R_1$ but in the second $E_2 \approx A L^{-1} R_1$ so the p_k factor will enter. This explains Fig. 5.

Wilkinson [12] presents the theory which shows that the eigenvalues of a matrix with larger p_k will grow faster with respect to perturbations of the matrix. This explains Table 3 since in our case the perturbation is h times the "b" coefficient matrix. Increased stability will be obtained with smaller p_k .

The matrix approach and particular the error bound with p_k factor explain very well all the graphs and tables containing the observed results.

4.3 Error in the First Cycle:

To understand more fully the observed behavior we must study the first cycle in more detail than an error bound will allow. At $h > 0$ the X_k are independent so E_0 and B_1 can be expanded as

$E_0 = \sum_{k=1}^{K} e_k X_k$ and $B_1 = \sum_{k=1}^{K} b_k X_k$. Multiplying by Z_k^T and using the definition of p_k gives $e_k = p_k E_0 Z_k^T$, $b_k = p_k B_1 Z_k^T$, and $E_1 = \sum_{k=1}^{K} p_k [\lambda_k E_0 Z_k^T + B_1 Z_k^T] X_k$.

A similar expression could be obtained for E_S from (4.2.6) in which λ_k^S would appear implying that only the principal condition numbers matter . It may be possible to pick E_0 such that the expression in brackets is 0 . In the general case there are two difficulties with this approach: (i) if truncation error $\leq$ roundoff error the second term is random and there is no hope of estimating it, (ii) if truncation $>>$ roundoff error the second term will be extremely difficult to estimate requiring knowledge of the higher derivatives of the solution. If this is the case it may be possible to go to a more accurate method at the same h (higher order, smaller $C_i \neq 0$, or a completely different method) . The "cancellation" effect (4.1) in the cyclic methods could be taken advantage of in either (i) or (ii) .

Supposing the local errors are $<<$ E_0; $E_S = \sum_{k=1}^{K} p_k \lambda_k^S E_0 Z_k^T X_k$. This also represents the propogated effect of the local error at a single step. For simplicity we will study $E_0 = (0, 0, 0, 0, 1)$ for the Cowell and cyclic $K = 5$ methods at $h = 10^{-4}$ using Tables 1, 3, and 4. The products of the extraneous $p_k \lambda_k$ were $< .01$ of the principle products even in the first cycle so they will be ignored. The first componant of E_1 for Cowell

$$= (6.326 \times 10^{-5})^{-1}(1.00050)(.70714)(.4471) + (-6.324 \times 10^{-5})^{-1}(0.99950)(.70707)(.4473)$$

$$= (1 + 5h)(.4471)(.70707 + .7h)/(.6324h)\left(1 + \frac{2}{.63} h\right) - (1 - 5h)(.4471 + 2h)(.70707)/(.6324h)$$

$$\approx (.4471)[.70707 + (5 \times .7 + .7)h + 3.5h^2][1 - \frac{2}{.63} h]/.6324h - (.70707)[.4471 + (2 - 5 \times .44)h - 10h^2]/.6324h$$

$$\approx [(.4)(.7)(-\frac{2}{.63}h) + .4(5x.7 + .7)h(1-\frac{2}{.63}h) - .7(2-5x.44)h]/.63h$$

≈ 7. We see that although $s_i \sim O(h)$ the fact that $|\lambda_1| - |\lambda_2| \sim |X_1| - |X_2| \sim |Z_1| - |Z_2| \sim O(h)$, that $|s_1| - |s_2| \sim O(h^2)$, and that there was an overall difference of sign all lead to the cancellation of the zeroth order term in the numerator thus making the error reasonable; a factor of 7 is in agreement with Table 1.

It may be possible to apply perturbation theory [12] to verify the above conditions (perhaps even the sign difference) on arbitrary $f(x, y)$, provided the p_k are not too large, since $\lambda_1 = \lambda_2$ and $X_1 = X_2$ at $h = 0$. The X and Z pairs must have nearly = corresponding componants.

For the cyclic method $|\lambda_1| - |\lambda_2| \sim |X_1| - |X_2| \sim |Z_1| - |Z_2| \sim O(h)$, $|s_1| - |s_2| \sim O(h^2)$, and there is a net difference in sign so also here

(4.3.1) $||E_1|| \sim c\ h\ ||E_0|| / |s_1|$ where c = constant + $O(h)$ includes a term $|s_1 + s_2| / |s_1| h$ so E_1 includes a term $\sim ||E_0||\ |s_1 + s_2| / |s_1|^2$. For the cyclic however

$||E_1|| \sim c\ 10^{-4}/10^{-8}\ ||E_0|| \sim ||E_0|| \times 10^4$ which is in surprising agreement with Table 1 in view of the simplifying assumptions made.

4.4 Optimization Criteria and Methods:

The problem facing us in not one of using some parameters to satisfy a set of linear or nonlinear equations as almost all procedures for deriving numerical methods are. We do not know what values can be

attained by criteria such as $||A||$ nor do we know how many parameters it would take to solve such an equation. The former problem is surmountable by just assuming smaller and smaller values from the present one. However, the wrong choice of number or type of parameters would make an equation solution impossible.

Although nonlinear optimization procedures are computationally more complex and lengthy than solving nonlinear equations they do solve the above problems. There are many ways to state our problem: (i) minimize $||A||$ subject to order and extraneous eigenvalue conditions, (ii) minimize extraneous eigenvalues subject to order and $||A||$ conditions , (iii) minimize the $|p_i|$ factors subject to order and eigenvalue conditions or vica versa. The condition of A or the norm of L or other ways of stating the primary error source could be included. Auxiliary conditions could include local truncation error, stability of the eigenvalues, the size of the coefficients themselves, and perhaps even some conditions on the behavior with respect to random local error (roundoff) . For $K = 3$ a simple mapping program was used to "optimize" certain criteria. This is too expensive a process if reasonable accuracy is required at $K = 3$ and for any accuracy at $K = 5$.

The iterative, nonlinear optimization algorithms seem to fall into three classes: (i) require no derivatives , (ii) require the gradient of the object function, and (iii) require second partials of the object function. Methods of (i) are slow converging and require as many function evaluations as (ii) [11] . In our problem we do not have an explicit expression for the object function much less its derivatives. Difference quotients have been used in (ii) but would be very inaccurate in (iii) [7] . The programming of methods (iii) is also complex. The (ii) methods, then, consist of the linearly convergent steepest descent (in the negative gradient direction) [1] and the quadratically convergent conjugate gradient [8] methods.

Several modifications of the latter are being tried under the program name OPTIMA. We are trying to reduce the total number of function evaluations since these will probably involve computation of eigenvalues and vectors which is very expensive. For example, to optimize a K = 5 method using 5 parameters will require $\geq$ 12 eigenvector computations per iteration for, perhaps, 200 iterations $\approx$ 2500 evaluations at about 1 second each at about \$.20 each second $\approx$ \$500; which does not satisfy budget constraints.

Once this program is working the constraints must be added. The best way to do this seems to be the penalty function method. Whenever the minimum search wanders outside the region where constraints are satisfied a penalty proportional to the size of the constraint is added to the function we are trying to minimize [6]. This tends to keep the search within the constraint boundaries.

The procedure will then be to solve the order equations parametrically, minimize $||A||$ subject to extraneous eigenvalues $\leq$ say $\frac{1}{10}$. We will then work on the more expensive minimize $|p_i|$ procedure which promises greater improvement. If imporvement up to Cowell is obtained then other optimization criteria can be added. If no improvement, order will be dropped to Cowell's and the above repeated. At this point, this work would begin to blend with optimization of traditional methods [5] so if no improvement is obtained over the already optimized traditional methods, then this phase of the work will be terminated.

5. OPTIMIZATION RESULTS

Using OPTK3 several K = 3 order 4 methods were derived that are both better and worse than Cowell with respect to $||A||$, $|s_1|$ and $|s_1 + s_2|$ at various h on $y'' = y$ (Table 5). The methods are not optimal due to the crudeness of the program, however improved computational accuracy is shown for some of the methods.

$||A||$ did not vary much over the h interval considered. Also $s_1 = 1/d_1$, the larger the better, and $|s_1 + s_2|$ should be $\sim O(h^2)$, the smaller the better. The integrations were done in double precision. The errors are the last componant at the first and 100th cycles. Method #7 is Cowell's.

Table 5 Some K = 3, Order 4 Methods

Method #	$\|\|A\|\|$	h	$\|s_1\|$	$\|s_1 + s_2\|$	error 1	error 100
1	9.0	0	4.0E-15	5.6E-26	5E-20	4E-17
		E-6	2.7E-7	1.7E-12	1E-27	1E-24
		E-4	2.6E-5	2.E-8	1E-25	6.1E-22
2	9.0	0	1.4E-14	9.2E-28	2E-19	2E-18
		E-6	1.2E-6	1.1E-11	6E-28	3E-25
		E-4			8E-26	6.2E-22
3	4.8	0	3.E-14	3.E-27	6E-20	1E-17
		E-6	6.8E-7	4.3E-12	2E-27	2E-23
		E-4	6.8E-5	4.3E-8	1E-25	6.1E-22
4	6.2	0	1.E-14	1.E-28	2E-19	3E-18
		E-6	1.1E-6	8.1E-12	2E-28	2E-25
		E-4	1.1E-4	8.1E-8	9E-26	6.3E-22
5	6.7	0	8.E-15	2.E-27	2E-19	2E-17
		E-6	1.0E-6	5.6E-14	2E-28	2E-24
		E-4	1.0E-5	5.6E-10	9E-26	6.5E-22
6	6.4	0	1.E-14	2.E-27	2E-19	1E-17
		E-6	1.1E-6	6.5E-13	7E-28	6E-24
		E-4	1.1E-4	6.5E-9	9E-26	6.4E-22
7	7.0	0	9.9E-15	1.6E-28	2E-19	2E-17
		E-6	8.2E-7	8.2E-13	3E-27	1E-23
Cowell		E-4	8.2E-5	8.2E-9	8E-26	6.2E-22

The spread between s_1 values is greatest at $h = 10^{-6}$ as is the spread in computation errors. Methods with larger s_1 are better. For example methods #2, 4 at $h = 10^{-6}$ have the largest s_i and the smallest errors being a factor of 50 better than Cowell at cycle 100. Method #1 is an exception having worse s_1 and $s_1 + s_2$ than Cowell but smaller error at $h = 10^{-6}$. Methods that are "better" at one h seem to maintain this at other h also. The computation errors at $h = 10^{-2}$ were all the same perhaps due to dominating truncation errors or lessening effect of larger s_1.

Eigen computations with methods #1, 4, 6, and 7 show extraneous values remain at 0 and the principal ones = $1 \pm 3h$ as expected. The componants of the vectors X_1, X_2, Z_1, and Z_2 change as O (h) with h so the analysis of (4.3.1) will apply.

These results imply improvement in the condition numbers can give smaller integration errors.

6. SOFTWARE DESCRIPTIONS

ELLIPSE:

Integrates two body orbits with Class II cyclic or traditional predictor-corrector methods. The major part of the program is the same as the GEOSTAR subroutine CSTEP documented in [2]. The program uses exact starting values and prints the error in each of the three position components by using fixed point iteration to solve Kepler's equation. It has been modified to correct more than once. Computations were in double precision (about 16 places) on the IBM 360.

COMPAR:

Integrates $y' = \alpha y$ with Class I cyclic or traditional correctors and $y'' = \alpha y$ for Class II. The starting values are taken backwards from $x = 1.2$ so that methods of all K start integrating at the same x value allowing direct comparisons in the first cycle. Starting values and the error at each step are found using the analytic

solutions. Given the coefficients a_i^m and b_i^m the stability matrix, A, is formed as in (4.2.5) depending on the Class and α. It is applied once each cycle to obtain the solution. It is possible to add a specified local error, B_s, in each cycle in addition to the normal roundoff and truncation errors. If $h = 0$ the process implied by (4.2.6) is undergone where E_0 and B_s can be specified. They were usually chosen to be (1, -2, 4, -8, 16) E-20 and 0 respectively, however, at times $B_s = E_0$. Computations were in single (about 14 places) or double (about 28 places) precision on the KRONOS time sharing system on the CDC 6400.

EIGENP:

Computes eigenvalues and vectors for each stability matrix, A, using the QR, double step, inverse iteration algorithm [12]. The Class II matrices at $h = 0$ were ill-conditioned so the program either did not work at all or gave poor accuracy. Program HESSEN (QRIEG) supplied by Mel Velez was used to verify (eigenvalues only) some cases. A later version of the program, EIGEN, also computes eigenvectors for A^T and the condition numbers. Computations were done in extended double precision (about 32 places) on the IBM 370 and in double precision (28) on the KRONOS system. An input parameter specifies the approximate number of places of desired accuracy before iteration terminates. We usually got by with 20 decimal places. Computing time for a 5 x 5 run with EIGEN $\approx$ 1 sec. $\approx$ \$.20 on KRONOS.

OPTK3

Computes $||A||$, max extraneous root, $|s_1|$, and $|s_1 + s_2|$ for a specified h for $y'' = \alpha y$ using the parametric order equation solutions at each grid point of an input specified grid for the free parameters. The program functions for Class I or II at any K or

order when the correct order equation solutions are given. Using the program in a man-machine, time sharing mode the user specifies the initial grid by typing in the left and right hand endpoints and step-size for each of the free parameters. After visually inspecting the output for the "best" region he immediately specifies a finer grid in this region and repeats, obtaining better methods. EIGEN is called at each grid point so this program was too expensive at K = 5 costing about $10 per run for very coarse grids. The KRONOS system was used.

OPTIMA:

Still being developed. Minimizes a nonlinear function of n variables, f, for which no explicit form is given. In the i th iteration one must choose a stepsize to approximate the partial derivatives of f, Δx, a downhill direction, D_i, and a step length in this direction, z. After considering several algorithms [1, 6, 7, 8, 11] it was decided to start with the method of steepest descent where $D_i = - \nabla f(X_i)$ and z was found by minimizing a parabola fit thru X_i, the directional derivative in the D_i direction (= slope of parabola), and a second, arbitrarily chosen point in the D_i direction. Per iteration, this algorithm required only n + 2 f evaluations. Convergence was good on simple quadratic f for small n but slow for the function with a narrow, curved, "banana" shaped valley [8]. For $f = ||A||$ for K = 5 , n = 10 with no constraints the parabola minimum overshot f min. so often that X_i was changing too radically to converge. The problem here is that $||A||$ is a high degree polynomial in the free parameters, too steep for a quadratic. Even if we could get down these steep walls we would be in a narrow, curving valley like the "banana".

The program was modified so that $D_i = - \nabla f(X_i) + D_{i-1} |\nabla f(X_i)|^2 / |\nabla f(X_{i-1})|^2$ called the conjugate gradient direction [8]. Theory states that if f min lies in a long, narrow,

quadratic valley then convergence is assured in $\leq n$ iterations while the steepest descent will take many more. For general functions one must set $D_i = -\nabla f$ every $n + 1$ iterations. This algorithm did converge faster on the banana but still overshot on $||A||$. Instead of a parabola fit, Fletcher uses a cubic fit to two points in the D_i direction and their directional derivatives. This requires $\geq 2n + 2$ f evaluations and possibly $2n+2$ more since he sometimes fits another cubic to obtain a better approximation to f min in the D_i direction. Our algorithm is being modified to successively fit parabolas to obtain a better approximation to f min in the D_i direction. We have taken $\Delta x = \frac{1}{2} z$ and this seems to work.

With the latest version of our algorithm after 50 iterations with $50 (2 \times 2) = 200$ f evaluations we converge to the same point on the banana as Fletcher does after 20 iterations with $\geq 20 (2x2 + 2) \geq 120$ f evaluations. His actual number of f evaluations could be as high as 200 but he does not give this data.

Part of the programming support for ELLIPSE and OPTIMA was cosponsored by the NASA-PSS contract, part of EIGENP by C. Shipp for graduate course credit at CSUF, and the remaining by E. Spiehler under the NASA-CSUF Grant.

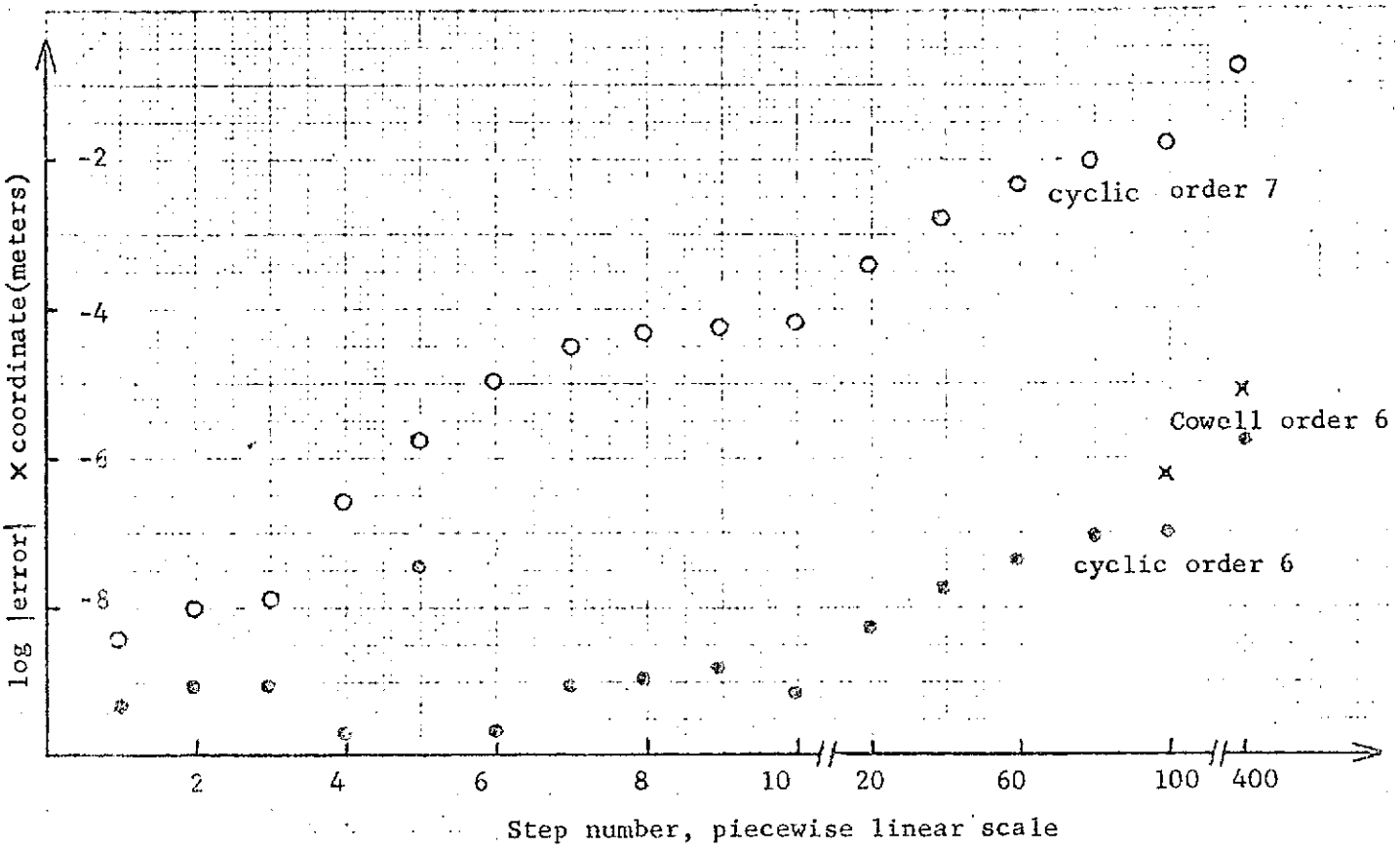

Figure 1. Cyclic vs. Cowell on an elliptic orbit at h=1 sec. for 400 secs. Roundoff error dominates.

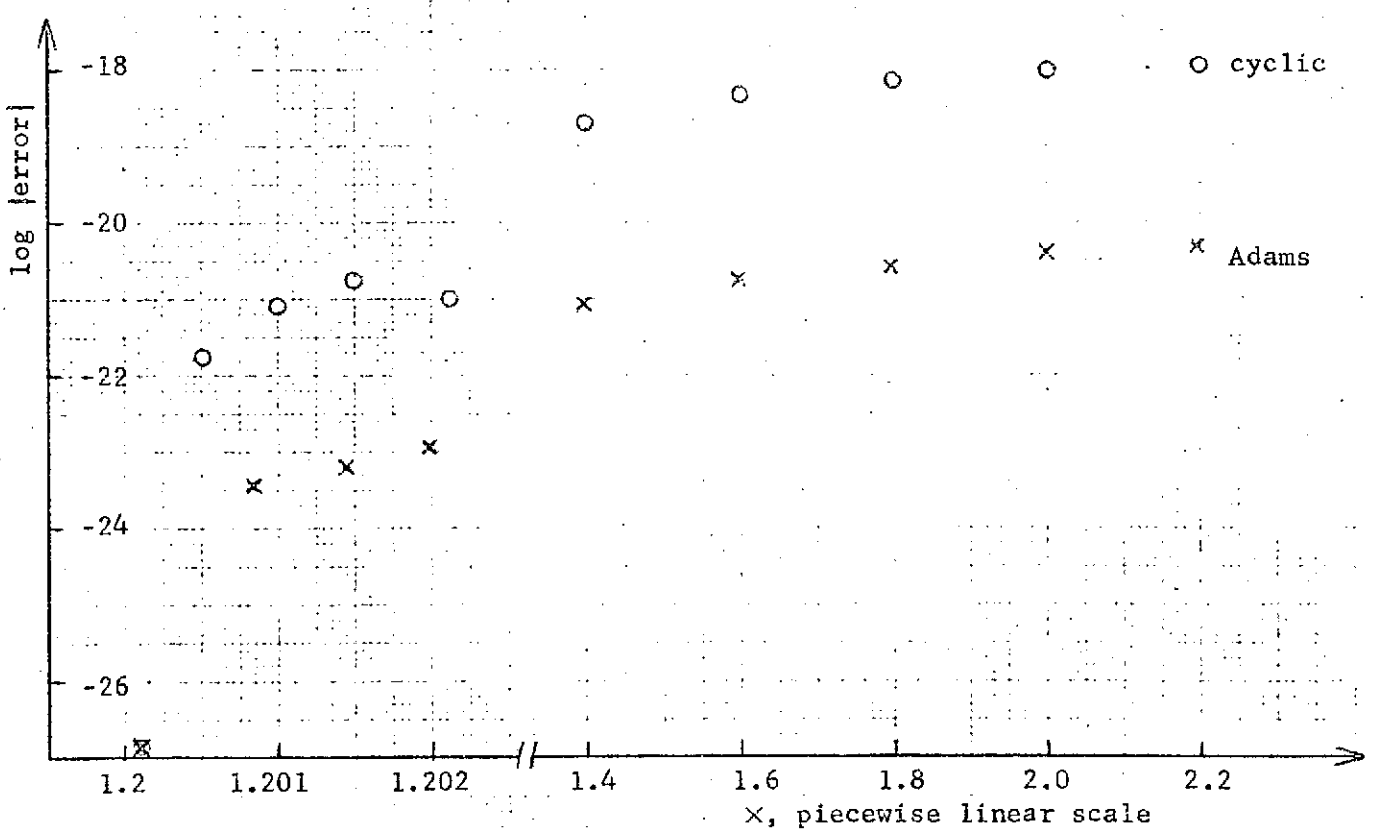

Figure 2. Cyclic vs. Adams on $y' = .5y$ at $h=10^{-4}$ for 2000 cycles. Roundoff error dominates.

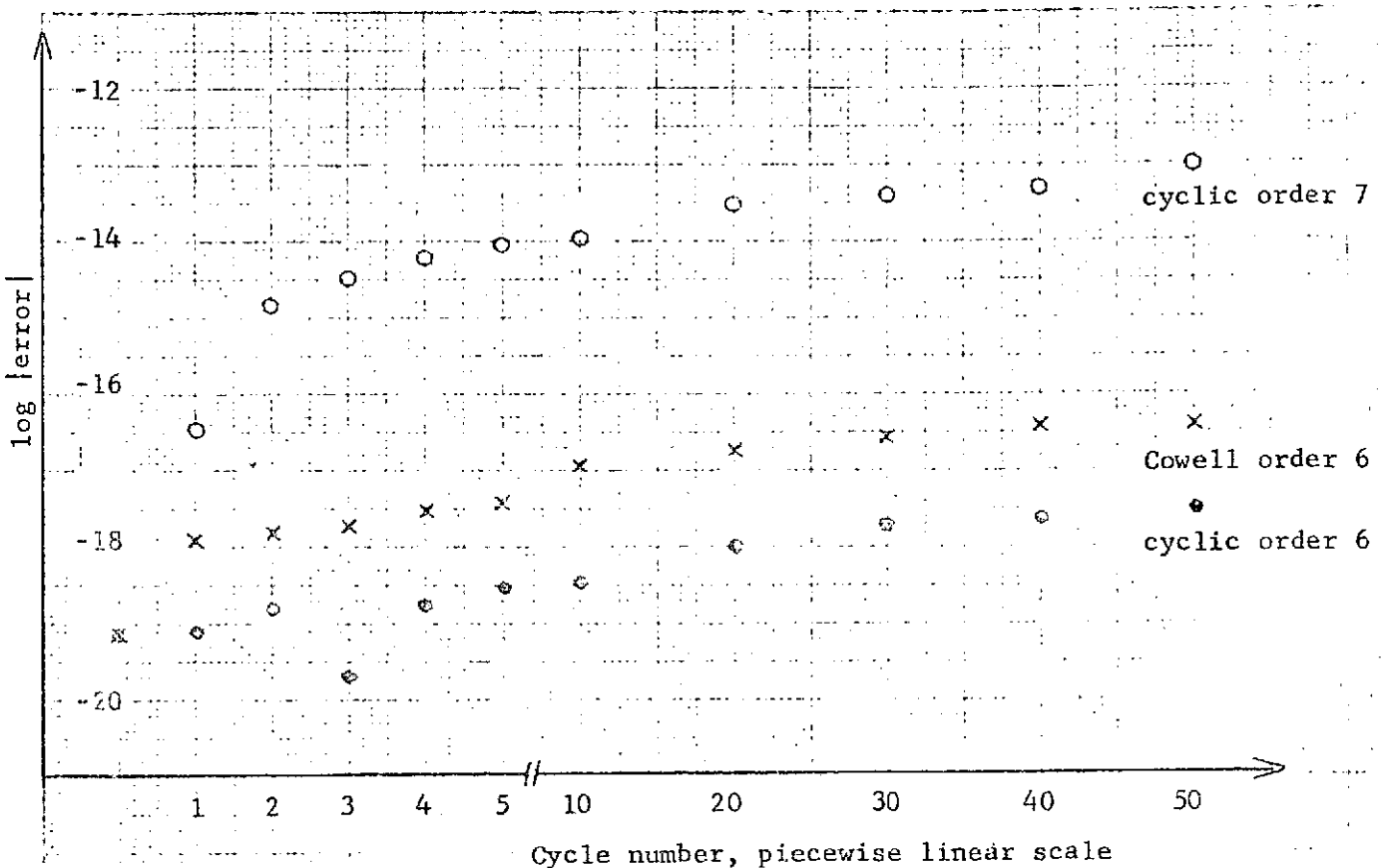

Figure 3. Cyclic vs. Cowell on starting errors only at h=0. Starting error dominates.

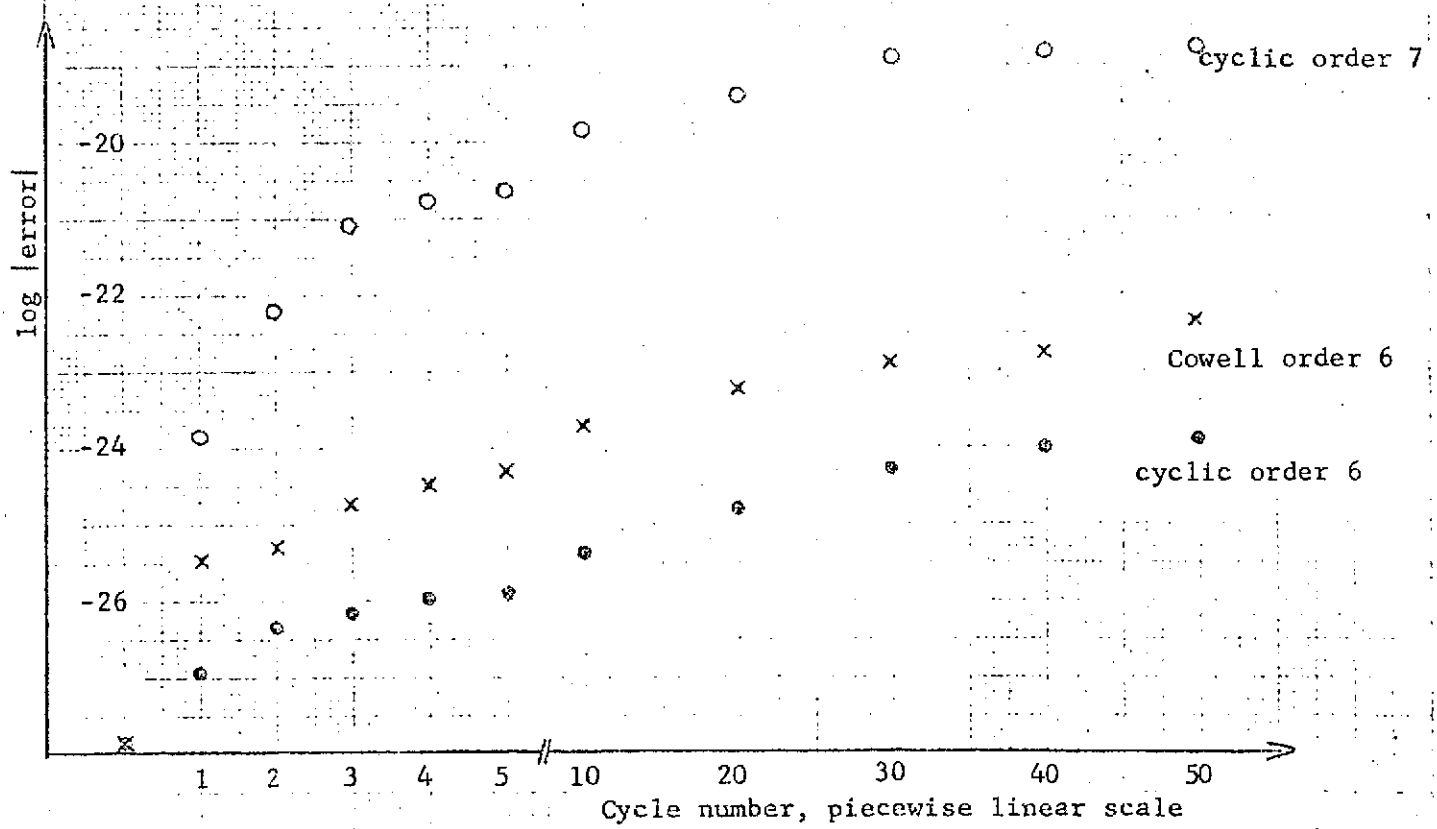

Figure 4. Cyclic vs. Cowell on $y'' = y$ at $h=10^{-6}$. Roundoff error dominates.

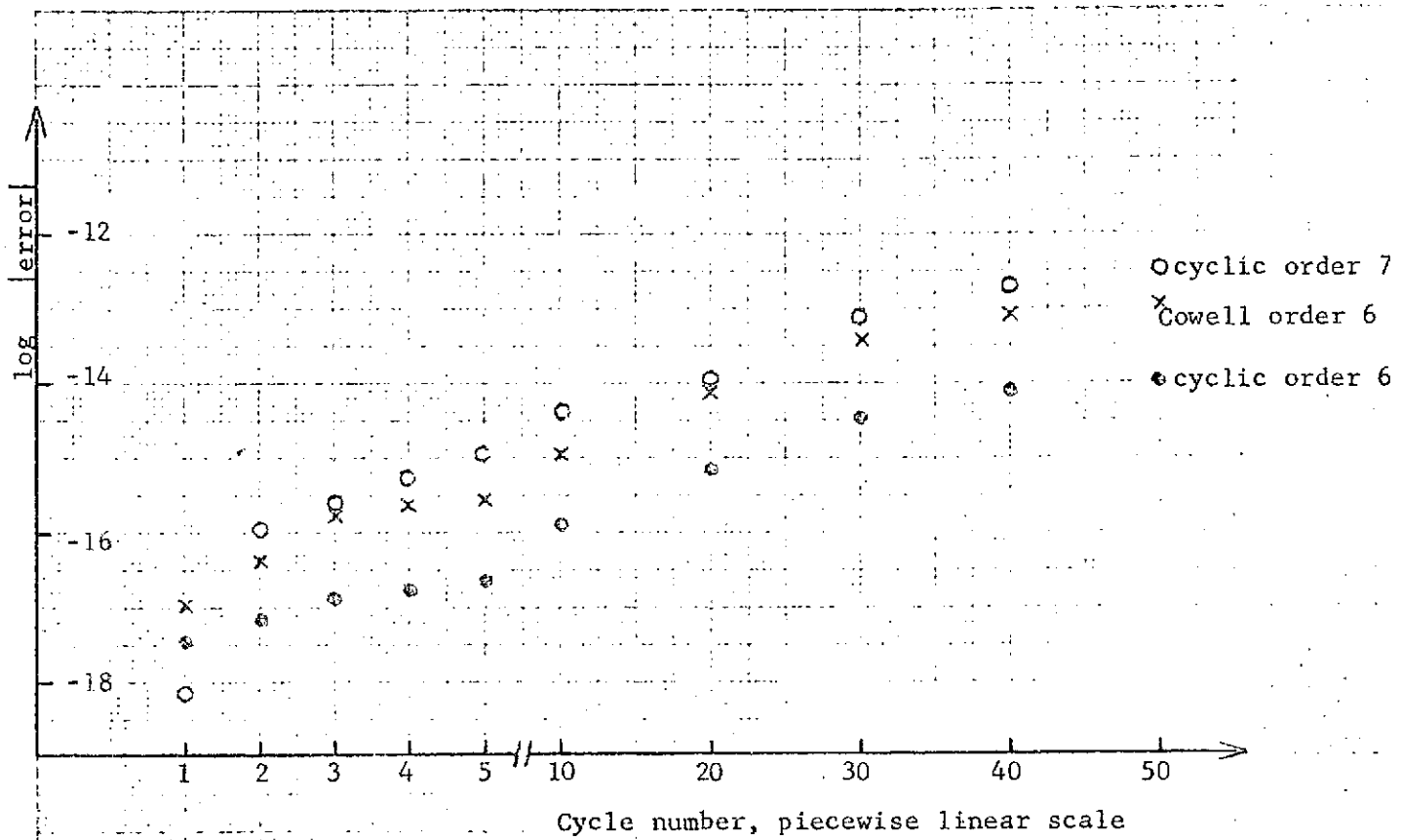

Figure 5. Cyclic vs. Cowell on $y'' = y$ at $h = 10^{-2}$. Truncation error dominates.

8. REFERENCES

1. Aoki, "Introduction to Optimization Techniques", Macmillan and Co., N.Y., (1971).

2. Chesler, L. and Pierce, S. "Application of Generalized, Cyclic and Modified Numerical Integration Algorithms to Problems of Satellite Orbit Computation", SDC TM-4717, Santa Monica, California, (1971).

3. Dahlquist, G. "Convergence and stability in the numerical integration of ordinary differential equations", Math. Scand. 4 (1956) 33-53.

4. Donelson, J. III and Hansen, E. "Cyclic composite multistep predictor-corrector methods", SIAM J. Num. Anal. 8 (1971) 137-157.

5. Dyer, J., Pierce, S., Haney, R. and Chesler, L. "Generalized Multistep Methods in Orbit Computation", SDC TM-4888, Santa Monica, California, (1972).

6. Fiacco and McCormack "Nonlinear Programming", Wiley, N.Y. (1968).

7. Fletcher, R. and Powell, M.J.D. "A rapidly convergent descent method for minimization", Comp. J. 6 (1963) 163.

8. Fletcher, R. and Reeves, C.M. "Function minimization by conjugate gradients", Comp. J. 7 (1964) 149-153 .

9. Henrici, P. "Discrete Variable Methods in Ordinary Differential Equations", Wiley, N.Y., (1962).

10. Pierce, S. and Chesler, L. "Cyclic correctors for solving Class I and II ordinary differential equations", SIAM Rev. 14 (1972).

11. Powell, M.J.D. "An efficient method for finding the minimum of a function of several variables without calculating derivatives", Comp. J. 7 (1964) 155-162.

12. Wilkinson, J. "The Algebraic Eigenvalue Problem", Oxford Univ. Press, Clarendon, N.Y. (1965).